LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 3 : 9 September 2003

Editor: M. S. Thirumalai, Ph.D.
Associate Editors: B. Mallikarjun, Ph.D.
Sam Mohanlal, Ph.D.
B. A. Sharada, Ph.D.

PRELIMINARIES TO THE PREPARATION OF A MACHINE AID TO TRANSLATE LINGUISTICS TEXTBOOKS IN ENGLISH INTO TAMIL

S. Rajendran, Ph.D. & S. Kamakshi, Ph.D.

---

1. PROLOGUE

The present paper is the profile of the Ph.D. dissertation submitted to Tamil University by Miss S. Kamakshi under the supervision of S. Rajendran.

The demand for teaching Linguistics in Tamil has made it mandatory to look for a tool which helps in translating English Text books in English into Tamil. So a project with an aim to prepare a Machine Translation Aid (MTA) to translate Linguistics texts in English into Tamil was visualized and Kamakshi was entrusted to test certain hypotheses and view point on MTA. This has culminated into a dissertation by Kamakshi (July, 2001). Here in this paper we try to portray the strategy thus evolved to translate English Text books of linguistics into Tamil by means of MTA.

2. OBJECTIVES

The project has the following objectives:

1. To understand the different machine translation models which are in vogue for selecting a feasible model.
2. To study the language of linguistics so that the domain specific features of the language of linguistics is thoroughly understood before proceeding to translate the linguistics texts by using machine.
3. To correlate the structure of the source language, English and target language, Tamil so that the transfer model adopted for translation can be successfully manipulated.
4. To prepare a prototype of Machine Translation Aid (MTA) to transfer linguistics texts in English into Tamil.

3. HYPOTHESES

It is proposed to test the following hypotheses in this study:

1. Though the preparation of a completely automated machine translation is difficult to achieve, it is feasible to achieve a human aided machine translation or machine aided human translation.
2. A successful preparation of a machine translation aid needs a thorough linguistic analysis aiming at unfolding the structure of the source language and target language complemented by a correlative study aiming at understanding the mechanism transferring the source language into target language.
3. By making the text of the target language domain specific, i.e. linguistics text, the menace of meaning interpretation in machine translation can be very much reduced.
4. The two powerful properties of human language are:
    i. Nearby words provide strong and consistent clues to the meaning of a target word, conditional on relative distance, order and syntactic relationship.
    ii. The meaning of a target word is highly consistent within any given document.
5. The lexical transfer approach is a transparent and modular model that facilitates evaluation at every step in the procedural mechanism.

4. PRELIMINARIES TO THE PREPARATION OF MT AID (MTA)

The preliminaries to prepare MTA can be listed as follows:

1. Understanding the architecture of English structure
2. Understanding the architecture of Tamil structure
3. Correlation of English Structure with that of Tamil to find out the salient commonness and differences
4. Listing of transfer rules
5. Studying of style of Linguistics with special reference to Chomsky's Aspects of theory of Syntax
6. Preparation of a bilingual transfer dictionary

5. UNDERSTANDING THE ARCHITECTURE OF ENGLISH STRUCTURE

Here the architecture of word building, phrase building, clause building and sentence building mechanisms of English has been studied with the aim to contrast it with that of Tamil. The word building mechanism of English has been studied so that the meaning of a word form in terms of lexeme and inflectional and derivational affixes can be correlated with that of Tamil to facilitate lexical transfer. The different patterns of major phrases such as noun phrase, verb phrase, adjectival phrase, adverbial phrase and prepositional phrase of English have been identified and depicted in structural terms to implement parsing technique. The different types of clause structures of both independent and dependent clauses have been thoroughly studied to correlate them with that of Tamil. All the sentential patterns have been identified. The different types of sentences and their word order have been studied, and there by identified the ways by which parsing can be performed on the English text and get the needed parsed trees for lexical and structural transfer.

6. UNDERSTANDING THE ARCHITECTURE OF TAMIL STRUCTURE

Here the architecture of sentence building mechanism of Tamil has been studied. The mechanism of building words has been explained to facilitate building of 'Tamil analysis dictionary' and 'Tamil

generation dictionary' that help in designing lexical transfer module. The items and the arrangement of them in major phrases such as noun phrase, verb phrase, adjectival phrase adverbial phrase and postpositional phrase have been studied with the eye on transfer mechanism. The mechanism of building clauses by coordination and subordination has been elaborately dealt with to facilitate structural transfer. Different types of Tamil sentences and their word order have been studied with the aim to transfer the English sentences into Tamil.

7. CORRELATION

Correlation of English Structure With That of Tamil to Find Out the Salient Commonness and Differences

Here the mechanism of transferring word forms, phrases, clauses and sentences in English into Tamil by correlating the structure of the two languages has been studied. The commonalties and differences in the structure of English and Tamil from the point of view of computation to build a machine translation aid to translate English into Tamil have been explored. The word building mechanisms of Tamil and English have been correlated with an aim to help in designing lexical transfer module. Noun phrases, verb phrases, adjectival phases, adverbial phrases and adpositional phrases have been correlated to understand the mechanism of transferring of these phrases in English into Tamil. It has been noticed that the two languages deviate from one another from the point of view of English as language of SVO (i.e. verb medial language) and Tamil as language of SOV (i.e. verb final language).

While English makes use of preposition to link nominal arguments with verbs, Tamil makes use of postposition and case markers to meet the same. The absence of regular case inflection in the case of English makes it rigid in its word order and the presence of case inflection in Tamil makes it more flexible in its word order. English distinguishes subject form object by means of the position, i.e. word order; where as Tamil does it by case inflection. Relative clause in English comes after the head noun, which is attributed and in Tamil it comes before the head noun. The infinitive clause in English comes after the main clause, whereas in Tamil it is comes before the main clause. That-clause complement occurs at the right side of

the main clause in English, whereas it occurs at the left side of the main clause in Tamil.

Interrogation is effected by changing the order of the words, i.e. by moving an auxiliary verb to the initial position before subject. In Tamil interrogation is effected by suffixing interrogative clitic or by making use of interrogative words. In English, the auxiliary verbs and the interrogative words occur in the initial position of the construction. In Tamil, the interrogative particles occur in the final position of any word in the construction. All these correlative features have to taken into account while designing the structural transfer module to restructure English as per Tamil sentential structure.

Under the head "Language of Linguistics" the characteristic features of language of Linguistics has been. The idea is to limit the machine translation system to work on a domain specific language so that the text to be translated can be kept under control. The salient features of language of linguistics have been studied so that the complexity involved in preparing an MTA for general use can be minimised. The unique characteristics of language of linguistics (i.e., the style of linguistics) have been identified. In terms of selection of vocabulary and technical terms linguistics differ from other disciplines. Its discourse structure is also unique in nature. It is a must for us to understand the language of linguistics, as the aim of the study is to translate linguistic textbooks in English into Tamil.

What is applicable to English is also applicable to Tamil as for as the language of linguistics is concerned. As Tamil is still a developing language as for as the linguistic discourse is concerned, the transfer of ideas form English to Tamil is a hard task. The problem of finding equivalents for technical terms coupled with transferring linguistic discourse in English into Tamil is a paramount. So it is essential to understand the unique features of language of linguistics before attempting to prepare an MTA for transferring linguistic knowledge in English into Tamil.

8. STRUCTURAL AND LEXICAL TRANSFER APPROACH FOR MT

A transfer model approach has been adopted to fulfill the aim of this study. Though it cannot be said that it is a full-fledged model, it definitely can convince one that an MTA prepared after correlative study of the source and target languages and by restricting the transfer to domain specific language (i.e. linguistics) will be a way ahead of other models. The step-by-step procedure of achieving the aim has been explained with illustrations. The preparation of a tagger and parser and a transfer model to transfer the linguistics texts in English into Tamil has been explained with the aim to lay a foundation for future research.

As the language differs in grammatical structure and lexical structure, one of the strategies to be adopted for the preparation of the MTA is to overcome these differences by altering the structure of the input to make it confirm to the rules of the target language. This can be done by applying contrastive knowledge, that is, knowledge about the differences between the two languages. Systems that use this strategy are sometimes said to be based on transfer model. It is proposed to adopt the transfer model for preparing MTA to translate linguistics texts in English into Tamil. Structural and Lexical Transfer (SALT) Approach coupled with PLNLP (Programming Language for Natural Language Processing) Approach explicated in Jesen et al (1993) is adopted to serve our purpose.

Any natural language processing system can be conceptually divided into three parts:

1. Grammar
2. Dictionary
3. The programming system that holds everything together

A division between grammar and dictionary (or lexicon) is inherited from linguistics, the study of language. The addition of a computational component is what turns this enterprise into a very current affair. Traditionally, grammars are systems of rules that mediate between symbols and meanings. The rules have a dynamic nature, and are supposed to embody generalizations that hold true for many symbols and combinations of symbols. Lexicons are repositories for particular units like words or phrases, and for information about those units.

Lexical information is prototypically static and specific in nature. In this context one may ask about the proper distribution of work between grammar rules and lexicon. Some systems based on versions of Chomskyan theory are to account for linguistic phenomena primarily in the rules.

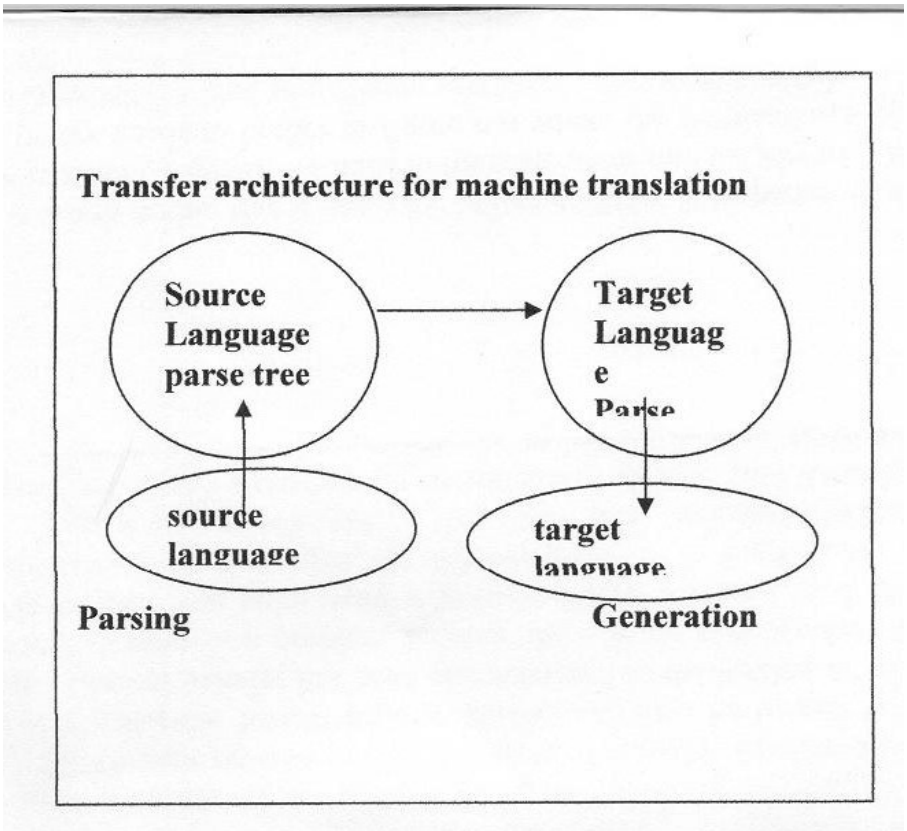
Another tendency, found typically in systems that derive from Lexical Functional Grammar, is to pack a lot of information into the lexical entries, and simplify the rules as much as possible. A certain amount of information is necessary to produce the analysis, and we have two poles for the distribution of that information: rules and lexicon. Different system chooses different ranges along the continuum.

There is an argument that the simplification of PS will complicate the lexicon and vice versa. There are no current standards of evaluation for assessing the relative merits of complexity of the lexicon versus complexity of the grammar rules, for the natural language processing systems. To the contrary, it seems likely that there is a certain amount of information that needs to be specified and manipulated, in order to parse any given sentence.

Sentence parser requires programming a computer to make a reasonable syntactic analysis of a sentence possible. Programming a computer to understand that sentence is more difficult, largely because the computer lacks the kind of prior experience that would enable to know what is supposed to understand. The analysis of a sentence using only syntactic information may contain many ambiguities. Attaching modifiers to their appropriate heads is the other main source of ambiguity. The strategy adopted for dealing with both kinds of ambiguity is that of packing the different syntactic descriptions into the same structure whenever possible.

Since the translation requires some representation of the structure of the input, transfer presupposes a parse of some form. Moreover, since transfer only results in a structure for the target language, it must be followed by generation phase to actually create the output sentence. It is worth noting that a parse for MT may differ from parses required for other purpose. For example, suppose we need to translate John saw the girl with binoculars into Tamil. The parser needs to bother to figure out where the prepositional phrase

attaches, because both possibilities lead to different sentences. So an MT system needs also to be able to represent disambiguated parses, while being able to work with ambiguities.



9. LEXICAL TRANSFER

By lexical transfer we mean the selection of the word, or words, in the target language that best render(s) the concept represented by the source language word or expression. The consequence of a particular lexical transfer choice will then be treated in the structural transfer phrase. Lexical transfer has been characterized as the bottleneck of MT systems. Basically, lexical transfer is a problem because there is no isomorphism among lexical fields across language. The same word may have several different (and often seemingly unrelated) translations. For example, for the noun subject in English has the meanings "kuTimakan", "peecappaTum viSayam", "ezuvaay" and "ndipandtanaikku uTpaTu" in Tamil.

This problem is very annoying for translation in general, and for MT in particular, because the correct choices can be conditioned by

circumstances ranging from general syntactic restriction to idiosyncratic properties at the lexical level. Moreover, the well-known problem of lexical gaps means that different languages choose different basic concepts to be realized as words and as phrase. To making things worse, even the way words interact in a single language does not seem to be a function of their meaning alone. And finally, there are idioms, frozen phrase that have lost their literal meaning and that simple to be memorized to be understood. Idioms are very culture-dependent, and more often than not cannot be literally translated. All of this raises doubt about the usefulness of word-to-word translation.

The foundation of lexical transfer is dictionary lookup in a cross language dictionary. The translation equivalent may be a single word or it may be a phrase, as in the example where gardening becomes tooTTaveelai ceytukoNTiru ('garden work doing'). Furthermore, sometimes a generation process must subsequently inflect words in such phrases, as in this case.

It is a well-known fact that a word may have several possible translations. Bank is such a word. It can mean 'a place where money is kept' and also 'the land near river, etc.' Fortunately there are at least two ways to tackle this problem: in the parsing or in the generation stage. The first method is to treat words like bank as if they were ambiguous. That is we assume that bank corresponds to two concepts and that choosing the correct Tamil word is like disambiguating between these concepts. This way of treating lexical transfer lets us to apply all the standard techniques for lexical disambiguation.

A second way is to treat such words as having only one meaning, and to handle the selection among multiple translations (karai 'bank of a river', vangki 'place where money is kept') by using constraints imposed by the target language during generation. In practice, these cases are more often dealt with in the parsing stage, as the algorithm for lexical choice during generation are high-overhead, especially for content words. In the case of examples of the following type the choice is easy.

He puts his entire savings in the bank.

Because the words with which bank is collocated makes the choice vangki 'place where money is kept' instead of karai 'place near river'. Such inputs, where multiple source language words must be expressed with a single target language word, can be difficult to handle, requiring inference in the general case. But many such cases, including this one can be treated simply as idioms, with their own entries in the bilingual dictionary.

10. STRUCTURAL TRANSFER

Our view of structural transfer does not impose any particular set of distinct linguistic levels at which transfer takes place. Depending on the problem at hand, a transfer may occur at any level of analysis, reflecting differences ranging from morpho-syntactic to cultural ones.

Morphological Level:
scissors → kattiri
PLURAL → SINGULAR

Syntactic Level:
She is beautiful → avaL azhakaanavaL
NP+Aux+Adj NP+Adj-PNG

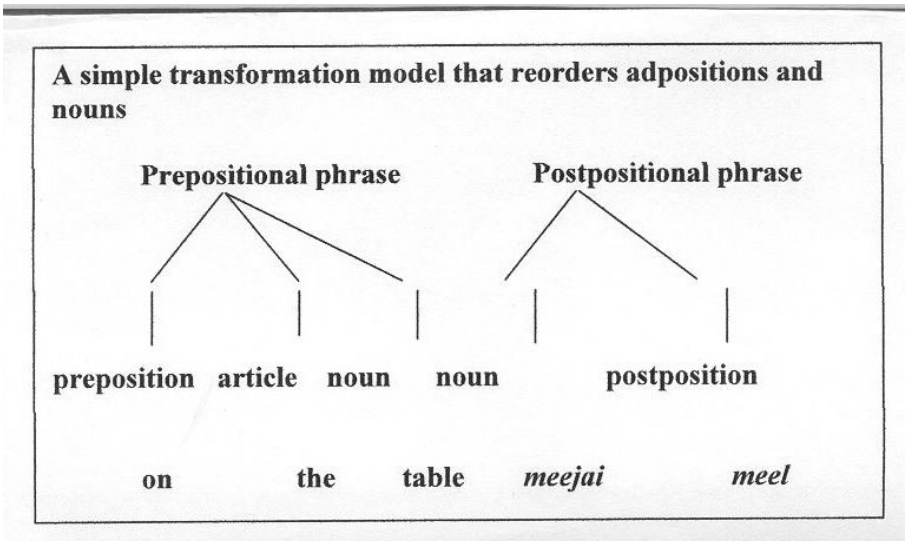
Cultural context level:
Please come in → vaarungkaL

Some of the output words are determined in the course of syntactic transfer or generation. In the example above, the function words are mostly grammatically controlled. Content words are another matter. The process of finding target language equivalents for the content words of the input, lexical transfer, is difficult for many reasons ranging from lexical conceptualization to grammatical conceptualization.

11. EXEMPLIFICATION ILLUSTRATING THE TRANSFER MECHANISMS

The use of postposition in Tamil in the place of preposition in English stands to exemplify the mechanisms of structural and lexical transfer. Temporarily postponing the question of how to translate the words, let's consider how MTA can overcome such

difference. (Adposition is a common term which is used here to cover up preposition and postposition.)

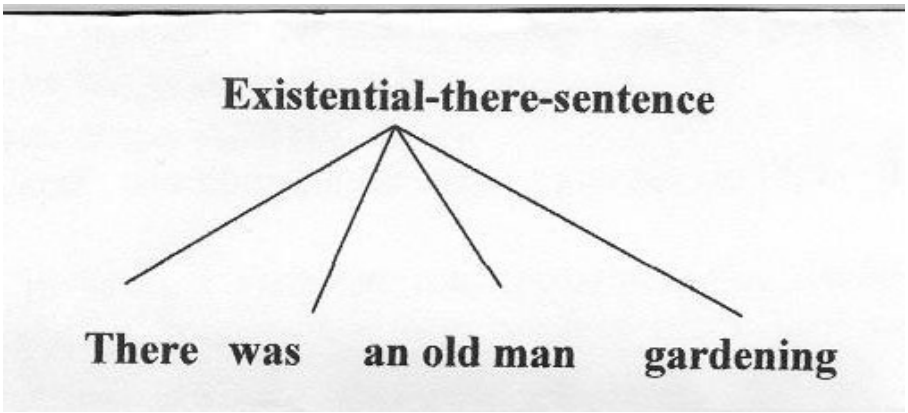


The above figure suggests the basic idea. Here one parse tree, suitable for describing an English parse, is transformed into another parse tree, suitable for describing a Tamil sentence. In general, syntactic transformations are operations that map from one tree to another.

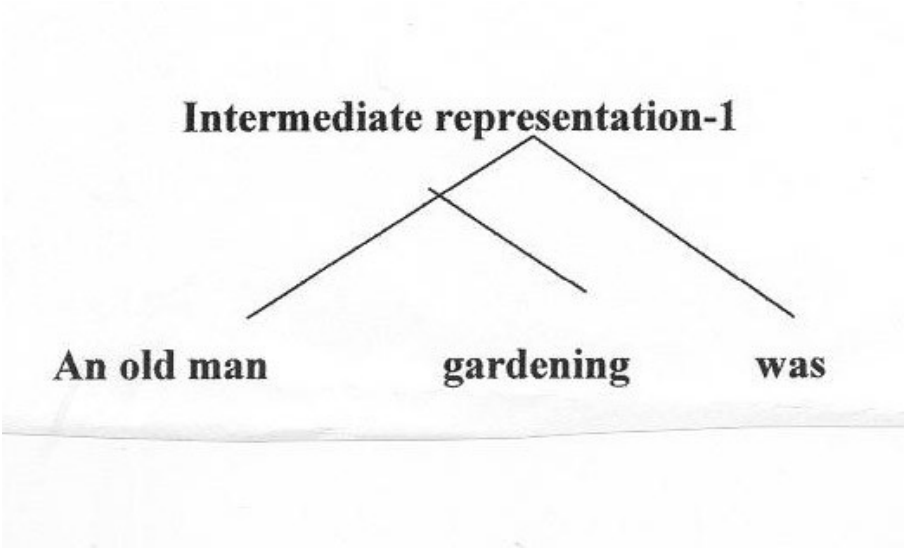
The following exemplification will illustrate roughly how such transformations can restructure an entire sentence.

There was an old man gardening.

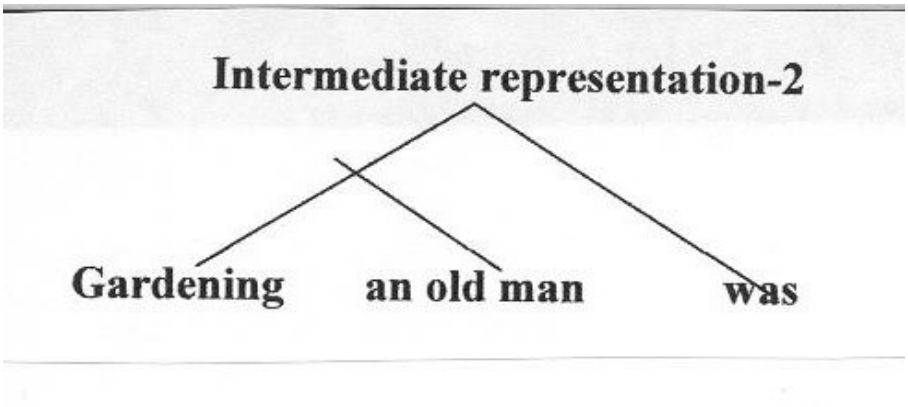
Assume that the parser has given a structure like the following:



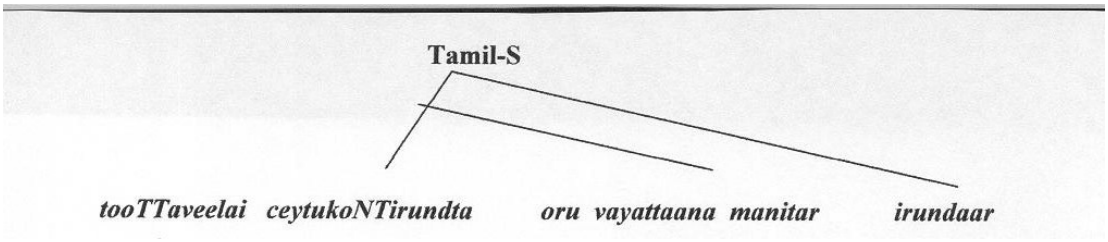
Since this sentence involves an "existential there construction", which has no analogue in Tamil, a transformation has to be applied to delete the sentence-initial there and convert the fourth constituent into a relative clause modifying the noun, producing something like the following structure:



The resulting structure is thus something more like the structure of a pseudo-English sentence: an old man, who was gardening, was. Next, another transformation applies to reverse the order of the noun phrase and the relative clause, giving something like the following structure:



At this point all relevant transformations have applied, and lexical transfer takes place, substituting Tamil words for the English ones. This gives the final structure given below:



A little more syntactic work is required to produce an actual Tamil sentence, including inflecting the verb. The final generation step transfers or otherwise linearizes the tree to produce a string of words. The final output will be:

tooTTaveelai ceytukoNTirundta oru vayataana manitar iruntdaar

The following table shows a rough representation of the transformations, which have been discussed above. Such transformation can be implemented as pattern-rewrite rules: if the input matches the left side of a transformation, it is rewritten according to the right side.

English to Tamil	
1.	PP → Preposition noun ⇒ PP Noun2 Postposition
2.	→ → Existential-There-Sentence There1 Verb2 NP3 Postnominal4 ⇒ Sentence (NP → NP3 Relative-Clause4) Verb2
3	NP → NP1 Relative-Clause2 ⇒ NP Relative-Clause2 NP1
→	

Table no.1 An informal description of some transformation

12. EPILOGUE

No claim is made here that a full-fledged automatic translation can be achieved by making use of the present prototype. But the prototype can be converted into a full-fledged package for translating linguistics texts in English into Tamil with human interference if proper changes are adapted to suit the new requirements. The exhaustive correlative study of English and

Tamil which amalgamates both contrastive and typological studies of the two languages will be a future documentation for one who will attempt for a full fledged MT system for transferring English into Tamil.

The scope of the present research work is a wider spectrum of identifying translation equivalents ranging from words to sentences and transferring the linguistic discourse in English into Tamil. The MTA can help a linguist who wants to translate a linguistics text in English into Tamil in many ways. It can help one to carry out off line translation tasks, the result of which can be available to the end users in the form of printed-paper books or E-books.

Select Bibliography

Asher, R.E. 1982. Tamil. Lingua Descriptive Studies. Amsterdam: North Holland Publishing Company.

Barathi, Akshar, Chaitanya, V. & Sangal, R. 1994. Natural Language Processing: A Paninian Perspective. New Delhi: Prentice-Hall of India.

Chomsky, 1965. Aspects of the Theory of Syntax. Cambridge: The MIT Press.

Cort, W. 1993. Typology and Universals. Cambridge: Cambridge University Press.

Crystal, D. and Davy, D. 1969. Investigating English Style. London: Longman.

Deshpande, W.R. 1994. Machine Translation: State of the Art (manuscript). Department of Electronics, Government of India.

Dakshinamurthy, R. 1983. A Contrastive Analysis of Complementation in English and Tamil. Ph.D. Thesis. Annamalainagar: Annamalai University.

Edward, W.T. 1973. A Contrastive Analysis of Tamil and English. Ph.D. Thesis. Annamalainagar: Annamalai Universtiy.

Geetha, K. 1985. Subsystems of Principles: A Study in Universal Based on Tamil Syntax. Ph.D. Thesis. Kanpur: IIT.

Gilman, W. 1961. The Language of Science. London: The English University Press.

Greenberg, et al. (eds.) 1978. Universals of Human Language. Vols. 1-4. Stanford: Stanford University Press.

Gopal, A. 1981. Adjectives in Tamil. Ph.D. thesis. Annamalainagar: Annamalai University.

Harris, Z. 1954. Transfer Grammar. In: International Journal of American Linguistics vol. 20, pp 259-270.

Hawkins. J.A. (ed.).1992. Explaining Language Universals. Oxford & Cambridge: Blackwell.

Hornby, A.S. 1975. Guide to Patterns and Usage in English. London: Oxford University Press.

Hutchins, W.G. 1986. Machine Translation: Past Present and Future. Ellis Horwood.

Hwee-Boon, L. 1993. Some Lessons Learnt by a New Comer (manuscript). Japan: MT Summit IV.

Isabelle, P. 1993. Machine-Aided Human Translation and the Paradigm Shift (manuscript). Japan: MT Summit IV.

Jensen, K, Heidron, G.E.& Richardson (ed.). 1993. Natural Language Processing: The PLNLP Approach. Boston/Dordrecht/London: Kluwer Academic Publishers.

Jurafsky, D. & Martin, J. H. 2000. Speech and Language Processing New Jersy: Prentice Hall.

King, M. (ed.) 1987. Machine Translation: The State of the Art. Edinburgh: Edinburgh University Press.

Kothandaraman, Pon. 1997. A Grammar of Contemporary Literary Tamil. Chennai: International Institute of Tamil Studies.

Leech, G.N. 1989. An A-Z of English grammar and usage. London: Edward Arnold.

Leech, G.N. et al. 1975. A Communicative grammar of English. London: Longman.

Lehmann, T. 1993. A grammar of modern Tamil. Pondicherry: Pondicherry Institute of Linguistics and Culture.

Lehmann, W.P. (ed.) Syntactic Typology: Studies in the Phenomenology of Language. Austin: University of Texas Press.

Newmark, P. 1988. A Text Book of Translation. New York: Prentice Hall.

Nida, E.A. 1974. Theory and Practice of Translation. Leiden: United Bible Societies.

------1975. Language Structure and Translation. Stanford, California: Stanford University Press.

Nirenberg, S. (ed.) 1987. Machine Translation. Cambridge: Cambridge University Press.

Palmer, F.R. 1986. Mood and Modality. Cambridge: Cambridge University Press.

Palmer, F.R. Grammatical Roles and Relations. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S. 1988. A University Grammar of English, London: ELBS, Longman.

Quirk, R., Greenbaum, S, Leech, G.N. and Svartvik, J. 1972. A Grammar of Contemporary English. London: Longman.

Quirk, R., Greenbaum, S., Leech, G.N. and Svartvik, J. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Quirk, R., Greenbaum, S. Leech, G.N. and Svartvik, J. 19 .C omprehensive Grammar of English.

Radhakrishnan. S. 1983. Noun Phrase in Tamil. Ph.D. Thesis. Annamalainagar: Annalmalainagar.

Rajendran, S. .Spell and Grammar Checker for Tamil. Paper read in Dravidian Linguists Conference. Trivandrum.

Rajendran, S. 1978. Syntax and Semantics of Tamil Verbs. Unpublished Ph.D. dissertation. Poona: University of Poona.

Rajendran, S. and Kamakshi, S. Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Texts in English into Tamil. Paper read in Dravidian Linguists Conference.

Ramachandran, S. 1975. In: PILC Journal Dravidic Studies.

Rangan, K. 1972. A Contrastive Analysis of the Grammatical Structures of Tamil and English. Unpublished Ph.D. Dissertation. Delhi: University of Delhi.

Rangan, K. & Chandrasekaran, G. 1994. A Glossary of standardized technical terms in Linguistics English-Tamil. Thanjavur: Tamil University.

Rangan, K., Suseela, M. & Rajendran, S. 1999. Some issues on the Relative Clause Constructions in Tamil (manuscript). Trivandrum: 27th All India Conference of Dravidian Linguistics.

Renuga Devi, V. 1997. Grammatical comparison of Tamil and English: A Typological Study. Madirai: Devi Publications.

Rohrer, C. 1993. The Future of MT Technology (manuscript). Japan: MT Summit IV.

Savory, T.H. 1967. Language of Science. London: Andre Deutsh.

Schiffmann, H. 1976. A Grammar of Spoken Tamil. Madras: CLS.

Shopen, T. (ed.). 1985. Language Typology and Syntactic description. vol i-iii. Cambridge: Cambridge University Press.

Sinha, B.K. 1986. Contrastive Analysis of English and Hindi Nominal Phrase. New Delhi: Bahri Publications.

Sivakumar, S. 1980. Complementation in Tamil. Ph.D. Thesis. Annamalainagar: Annamalai University.

Spencer, J. (ed.). 1965. Linguistics and Style. London: Oxford.

Subbaiah Pillai, K. 1983. A Contrastive Analysis of Interjections in Tamil and English. Ph.D. Thesis. Annamalainagar: Annamalai University.

Suseela, M. 1981. A Contrastive Analysis of Relative Clauses in Tamil and Hindi. Ph.D. Thesis. Annamalainagar: Annamalai University.

Syamala, V. 1992. A text book of English Phonetics and Structure for Indian Students. Trivandrum: Sharath Ganga Publication.

Syamala, V. 1994. A Textbook of English Phonetics and Contemporary Grammar. Trivandrum: Sharath Ganga Publication.

Theivanatham Pillai, K. 1970. A Comparative Study of the English and Tamil Auxiliary Verb Systems and Prediction of Learning Problems for Tamil Students of English. International Review of Applied Linguistics, vol 8, pp. 21-47.

--------. 1974. Contrastive Linguistics and Language Teaching. Annamalainagar: Annamalai University.

Thiyagarajan, K. 1881. Modal System of English and Tamil. Ph.D. Thesis. Madras: University of Madras.

Thirumalai, M. S. 1979. Language in Science. Mysore: Geetha Book House.